Data pre-processing for water resource modeling

E. S. Lillie[1], M. D. Watson[2]

[1] P D Naidoo and Associates, Consulting Engineers, Johannesburg, South Africa.
[2] Department of Water Affairs and Forestry, Pretoria, South Africa.

Abstract

From the analysis of the water resource modeling process, it was found that more than a third of all time spent goes into data collection and manipulation of input data. This results from the fact that data is not always very accessible and the manipulation of data is often cumbersome and time consuming. The automation of certain data pre-processing activities for the water resource modeling will reduce the time and cost of doing water resource evaluation projects.

The main objectives for the development of information management systems (IMS) to support data pre-processing for water resource modeling are to improve the efficiency and effectiveness of accessing, updating, verifying and sharing decision support data.

To achieve these objectives well-structured and easily accessible information management systems are required. These systems shall interface with the main source datasets, pre-process the data, perform quality controls and create the input information for water resource models.

The tools to manage the interface to the source data should be model independent. They shall keep metadata associated with any changes to the base data. This will allow for changes to be effectively shared with other modelers. The tools should allow for quick and efficient updates to the decision support data using the source data. The system will allow for models using different temporal resolutions, different spatial representations, different units of measurement and different modeling concepts. The systems will link to an independent GIS utility.

Each water resource model will require its own model specific pre-processor to configure the data into the correct format. The pre-processor shall interface with the appropriate information management system.

To develop all the pre-processors envisaged is a very significant undertaking which will take several thousand man hours to achieve. Therefore the pre-processors will be developed in stages. Each stage must also be usable as a stand alone system, however should also be designed to be able to integrate with other pre-processors in a single system at a later stage.

The creation of integrated information management systems for the pre-processing of decision support data will provide a platform for more efficient sharing of information.

Keywords: *pre-processor, information management, water resource modeling, Rain IMS*

1 Introduction

The purpose of this paper is to provide a vision for data pre-processors within an integrated analysis framework and to illustrate the current achievements made towards this vision. The pre-processors form key components of a water resources analysis framework. The pre-processors interface with the main source datasets, pre-process data for analysis purposes, perform quality controls and create input information for water resource models.

The main objectives for the development of pre-processors for water resource modeling are to:

- guide individuals on the various methods used for pre-processing data,
- improve the efficiency and effectiveness of accessing, updating and verifying analysis data,
- reduce the risk of knowledge loss,
- reduce duplicated effort of data acquisition and manipulation, and
- facilitate the sharing and communication of planning information for joint decision making and stakeholder participation.

2 Overview of Business Needs and Processes

The need for decentralised management of water resources is promoted through the National Water ACT (NWA). Improved information management for modeling is essential to the success of the decentralised management, communication and information sharing in the water resource planning process.

The high level flow of information in a typical water resource assessment is shown in figure 1 below. The pre-processing and the quality control of data is a critical step in the water resource evaluation process and needs to be performed with due care to ensure the integrity of the evaluation process.

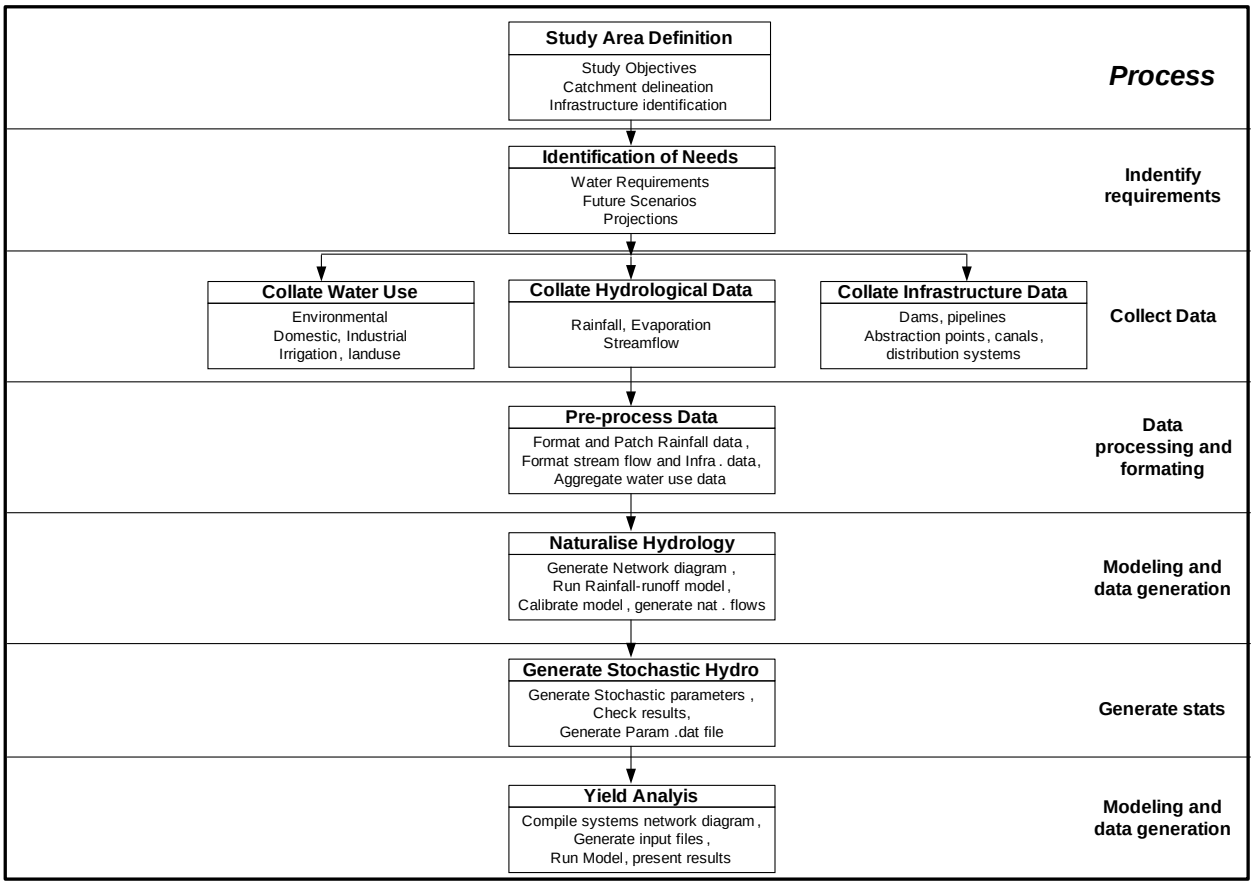

Figure 1: Water Resource Evaluation Process *(ref1)*

In current modeling practice, data is obtained from various sources for a model. The data is then pre-processed to be suitable as input for the specific model. The model is used to perform an analysis of the data. The output data is post processed for evaluation or alternatively to be used as input to another model. Three of the main information management short comings of this process are:

1) no seamless connection between the source data and the models
2) no consistency in the techniques used for preprocessing data between various organizations
3) poor metadata on the processes used to manipulate the source data to make it suitable as input data to a model.

This makes it difficult to share the processed data with other users or to easily check the assumptions used to pre-process the data. Often this results in duplicated effort in adjusting source data, because of the difficulties associated with assessing what has previously been done to obtain model input data. If an organisation is not exactly sure what processing the data has been subjected to, then it is very difficult for them to use the data with confidence or to take responsibility for its application.

3 Pre-processors information systems

There is considerable benefit in sharing well designed pre-processors IMSs for a range of raw/source data sets. This would facilitate the tracking of processes used to manipulate source data through the use of metadata. The source and processed data should always be stored and managed separately. It is envisaged that the pre-processor information systems will be embedded in an analysis framework for processed information. The framework has four main areas of functionality:

- warehousing utilities,
- pre-processing information systems,
- analysis systems and interfaces, and
- framework utilities.

Figure 2 below provides an overview of an envisaged analysis framework. This paper concentrates on the pre-processing information systems, however, it does provide some information on the other components.

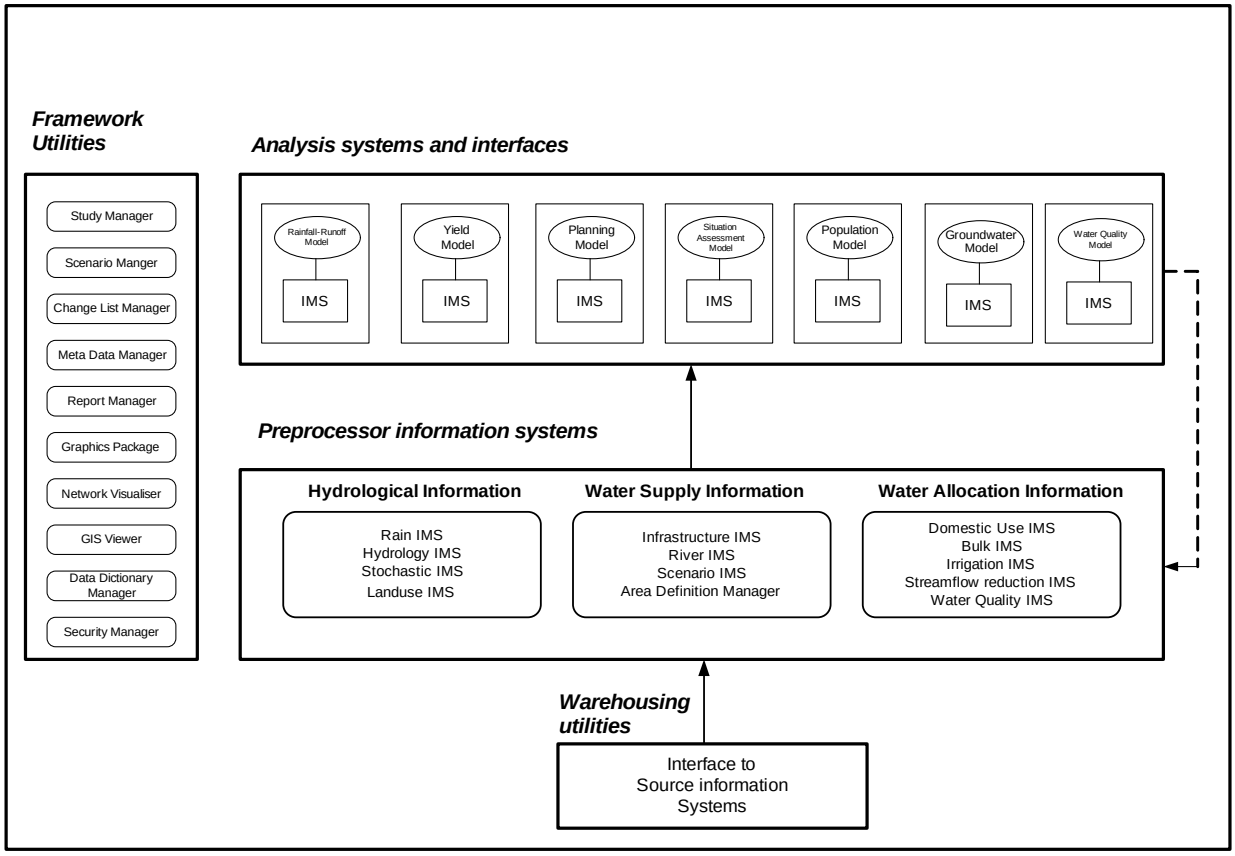

Figure 2: Possible water resources analysis framework

4.1 Warehousing utilities

The warehousing utilities manage the interface to the source data and should be model independent. The warehousing utilities allow for quick and efficient updates of source data at user defined intervals. They allow for models using different temporal resolutions, different spatial representations, different units of measurement and different modeling concepts.

Source information is defined as the "raw" data contained in databases that are continuously updated and maintained by various organisations, such as the South African Weather Services (SAWS) rainfall data, the measured streamflow records contained in the Department of Water Affairs and Forestry's (DWAF) hydrological data management system (HYDSTRA) and the water use information contained in DWAF's Water Administration and Registration Management System (WARMS). The management and maintenance of the source data is out of the scope of the envisaged analysis framework.

4.2 Pre-processors

The pre-processors manage the source data from warehousing utilities before it is incorporated into any specific model's information management system. The pre-processors allow for adjustments to the raw/source data while keeping track of all changes made. They perform quality controls on the source data with the aid of customised utilities. The source and modified data are kept in separate data structures, thus allowing subsequent users to access the original source data and acquaint themselves quality controls performed on the data. The pre-processors retain metadata associated with any changes to the raw/source data. This allows for any changes to be effectively shared with other modelers. It allows for easy and efficient checking of the assumptions and processes used to modify the raw/source data.

Each model used for analysis has its own unique data requirements and formats. However there is a huge overlap in terms of the type of data required by the different models used for water resource planning. A typical example is both the Water Resources Yield Model *(ref 2)* and the Pitman Model *(ref 3)* require information on the physical characteristics of a reservoir basin. However the same data is needed in different formats for each model. The Pitman Model requires the storage area relationship in terms of equation coefficients, where as the Yield Model requires the relationship in a tabular format, however both sets of data will come from the same source data. The pre-processors serve to configure the same raw/source (or modified raw/source data) into the appropriate format for any specific model supported by the system. This ensures greater consistency in the use of source information and the efficient application of added knowledge.

The pre-processors have for convenience been grouped into three main categories, hydrological information, water supply information and water allocation information. To develop all the pre-processors envisaged is a very significant undertaking which will take several thousand person hours to achieve. Therefore the pre-processors are being developed in stages, with those that can provide maximum immediate benefit being developed first. Each system is designed as an independently-usable stand-alone module. A whole range of data pre-processors have been identified for development. The first pre-processor to be developed is the Rain IMS, the next two pre-processors identified for development are a stochastic parameter IMS and a hydrology IMS.

4.3 Analysis systems and interfaces

Analysis systems and interfaces include water resources models with user interfaces and databases catering for each model's specific data requirements. The interfaces have functionality to validate the input data for each model and to share model configurations with other users.

The main objective of the user interfaces is to provide well structured information management. The interfaces are designed to obtain, store, access and process the information required for a specific analysis system/model. The interfaces also assist with the running of the analysis systems. The list of systems in Figure 2 is not exhaustive and the framework is intended to support any analysis system.

The first analysis system to be added to the framework is the Water Resources Yield Model (WRYM).

4.4 Framework utilities

The framework utilities are available to both the pre-processors and the models. There is a wide range of utilities that can be made available within the framework. The description of a few of these utilities is provided for illustration purposes.

The current framework utilities include a GIS utility to view both modeling and pre-processing data spatially. A network visualiser enables the modeler to view or build a network diagram for a model. A study manager is available to store and access information in the framework across models for a particular study/project. Change lists allow the modeler to make changes to the raw/source data, while keeping the original data intact. This makes it easier to manage changes to both existing source data and existing model configurations. A time series comparator is available to compare and view time series data.

5 Example of a pre-processor (Rain IMS)

The first pre-processor developed within the analysis framework is the Rain IMS, which allows for the selection, visualisation, quality control and finally the export of rainfall records for use in various models. Included in the Rain IMS are the systems ClassR and PatchR *(ref 4)*, which are used for the selection of rain gauges and the patching of missing or unreliable records.

5.1 Access to Data

The Rain IMS gives hydrological modelers access to the monthly rainfall records stored in the Hydstra database at DWAF as well as the Water Research Commission Project No. 1156 *(ref 5)*. This has the advantage of allowing modelers to work on a common dataset. Rain gauges may be selected using a variety of techniques such as a GIS facility, or by the SAWS number, or by buffering from a point.

5.2 Quality Controls

The Rain IMS has functionality to perform various quality controls on existing data sets to assist the modeler search the dataset for anomalous values. The search methods used to detect outliers include the tabulation of monthly and annual statistics, cumulative sum plots and cluster analysis. If gross irregularities are not detected in the rainfall records, these can have a significant impact on the hydrological modeling. There is also a time series comparator available to the preprocessor for consistency checking.

5.3 Modifying Raw/Source Data

The Rain IMS has the functionality to make changes to the raw/source data through changelists. These changelists make it easier for users to share modifications to the data and to allow for efficient audits on changes made to the raw/source data. The Rain IMS works on the basic principal that users may not physically change the raw/source data, they may only modify the data through a changelists which create an alternative value for the data while retaining the original record. In other words raw or source data is always stored separately from manipulated data for users to easily track modifications and the associated reasons. Changelists can be shared with other users.

5.4 Patching and infilling rainfall data

The Rain IMS has an interface to the systems ClassR and PatchR for patching and infilling of rainfall records. The Rain IMS maintains a record of the target and source gauges used in an analysis. The database also keeps a record of the input and output files used in the analysis.

5.5 Exporting Rainfall data for modelling

The Rain IMS can export any set of modified raw/source data to file format the former Hydrological Research Unit (HRU), through an export facility. This file format is the format required by many of the models being used for water resource modelling.

5.6 Sharing Information

There is an import and export facility in the Rain IMS which allows the users to archive and/or share the processed rainfall data with other users.

5.7 Reporting

There are fairly flexible methods to obtain information from the Rain IMS for inclusion in other documentation. The reporting facilities include the functionality to create sub-area reports, print graphs, export maps showing locality of rainfall stations and export rainfall records to spreadsheets.

6 Conclusions

Well-structured and easily accessible pre-processor information management systems are of great benefit to the water resource modeling community. The pre-processors serve to interface with the main source datasets through a set of warehousing utilities in order to extract the latest data at regular intervals. The pre-processors can effectively include methods to perform quality controls on the source data. The pre-processors serve to configure the raw/source (or modified raw/source data) into the appropriate format for a specific model supported by the framework. This will ensure that all modelers have access to the same raw/source data to ensure relative consistency between analyses. The pre-processors allow the user to efficiently check any changes made to raw/source data. The Rain IMS is the first preprocessor to be developed. The Rain IMS serves as an example of the concepts being adopted for the development of the preprocessors.

The ongoing development of pre-processor information management systems will provide a platform for more efficient sharing of information and will also improve the efficiency of accessing, updating and verifying data for water resource modeling. Through this the pre-processors will promote commonality in planning and facilitate effective and transparent decision making. New users will be given greater access to information and thus be empowered to participate more effectively.

References

1. Nyland G. 2002. Water Resource Evaluation Business Process Analysis. (Internal DWAF document)
2. McKenzie R. 1999. Water Resources Yield Model (WRYM): User Guide- release 4.1.1
3. Pitman W, Kakebeeke J and Bailey A. 2000. Water Resources Simulation Model for Windows (WRSM2000): Users Guide
4. Pegram G. 1997. Patching Rainfall Records-A Guide. (Internal DWAF document)
5. Lynch S, 2003. The Development of an Improved Gridded Database of Annual, Monthly and Daily Rainfall. Water Research Commission Project No. 1156